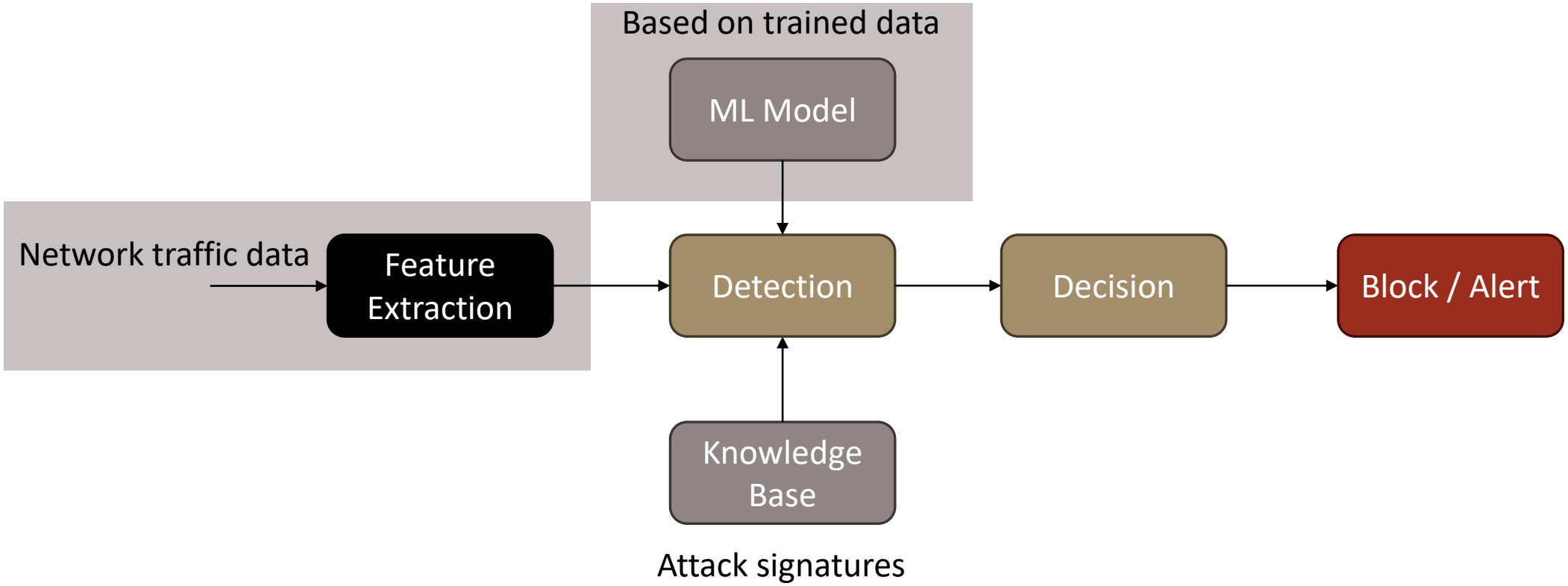
CSCE 585 Project Milestone
Intelligent In-Network Attack Detection

Sergio Elizalde, Amith GSPN, Samia Choueiri

Department of Integrated Information Technology
University of South Carolina

September 30, 2025

System Architecture



Data

Datasets — Canadian Institute for Cybersecurity
CSE-CIC-IDS2018-AWS, CICIDS2017, CIC DoS dataset(2016)

Datapoints — 12,794,627 (attack + benign)

Features — 84 features

	Src Port	Dst Port	Protocol	Flow Duration	Tot Fwd Pkts	Tot Bwd Pkts	TotLen Fwd Pkts	TotLen Bwd Pkts
mean	37081.266681	14638.242458	7.828970	8213385.375572	27.490721	4.727698	1142.314171	2719.337494
std	25217.602217	23063.516684	4.205032	24754440.702548	1741.563659	105.853364	55745.670576	152641.167058
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	443.000000	80.000000	6.000000	1263.000000	1.000000	1.000000	0.000000	0.000000
50%	50596.000000	80.000000	6.000000	32171.000000	2.000000	1.000000	42.000000	113.000000
75%	56222.000000	38510.000000	6.000000	4156602.500000	4.000000	4.000000	935.000000	358.000000
max	65535.000000	65534.000000	17.000000	119999996.000000	309628.000000	21676.000000	9908096.000000	31142967.000000

...

Data Balance

Datapoints

12,794,627 (attack + benign)

Limitations

Features

84 features

Features reduction: Calculating 84 features will

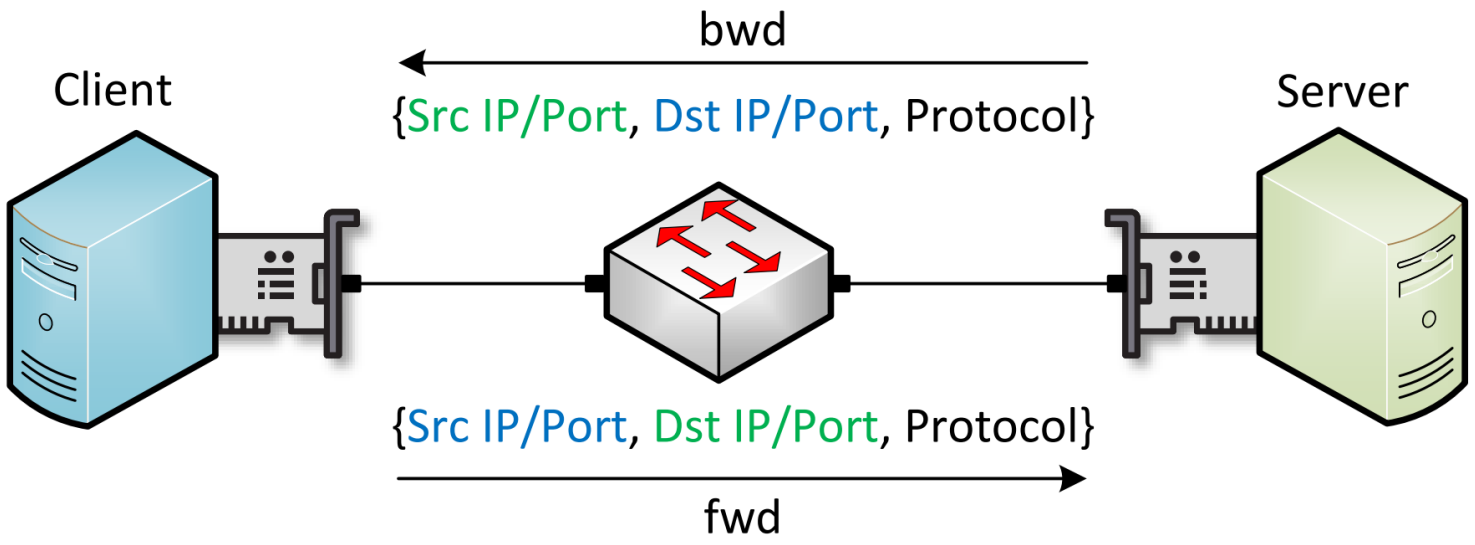
- Require high computational cost
- Induce latency and reduce performance
- Increase storage and memory requirements
- Lead to overfitting

Data Sanitizing

Data Sanitizing

Knowledge-based filtering | Correlation analysis (fwd) | Correlation analysis (bwd) | Final correlation analysis

- Network packets are grouped into flows.
- Flow ID is composed by 5-tuple: Src IP, Dst IP, Src Port, Dst Port and protocol.
- A session between client a server has two flows: forward and backward.
- Example:

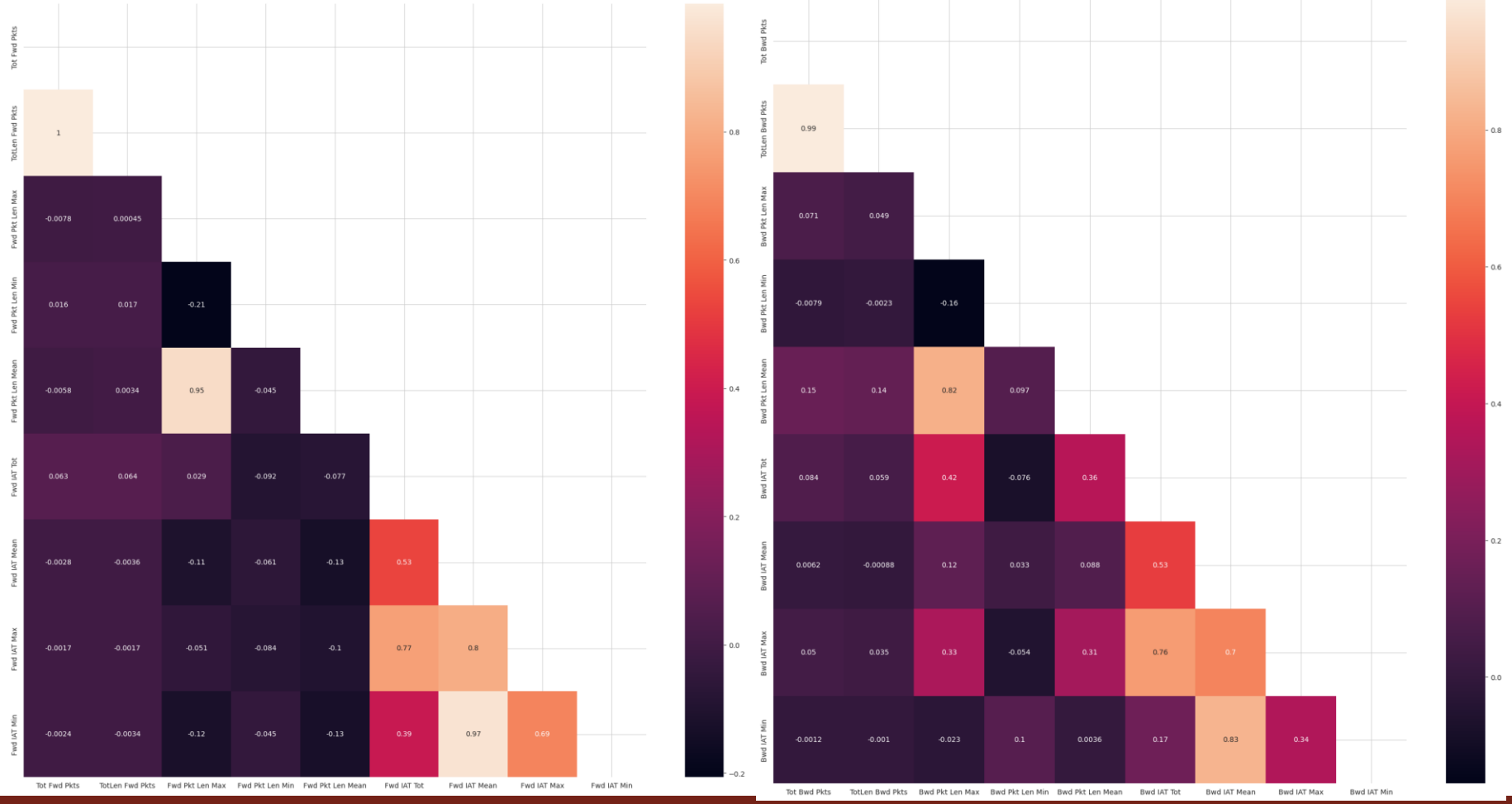


Data Sanitizing

- Level 1: Ambiguous / Redundant
- Level 2: High-Cost Statistics
- Level 3: Derived Features
- Level 4: Structural / Transport-Specific and Flow Information
- Level 5: Irrelevant / Low Utility

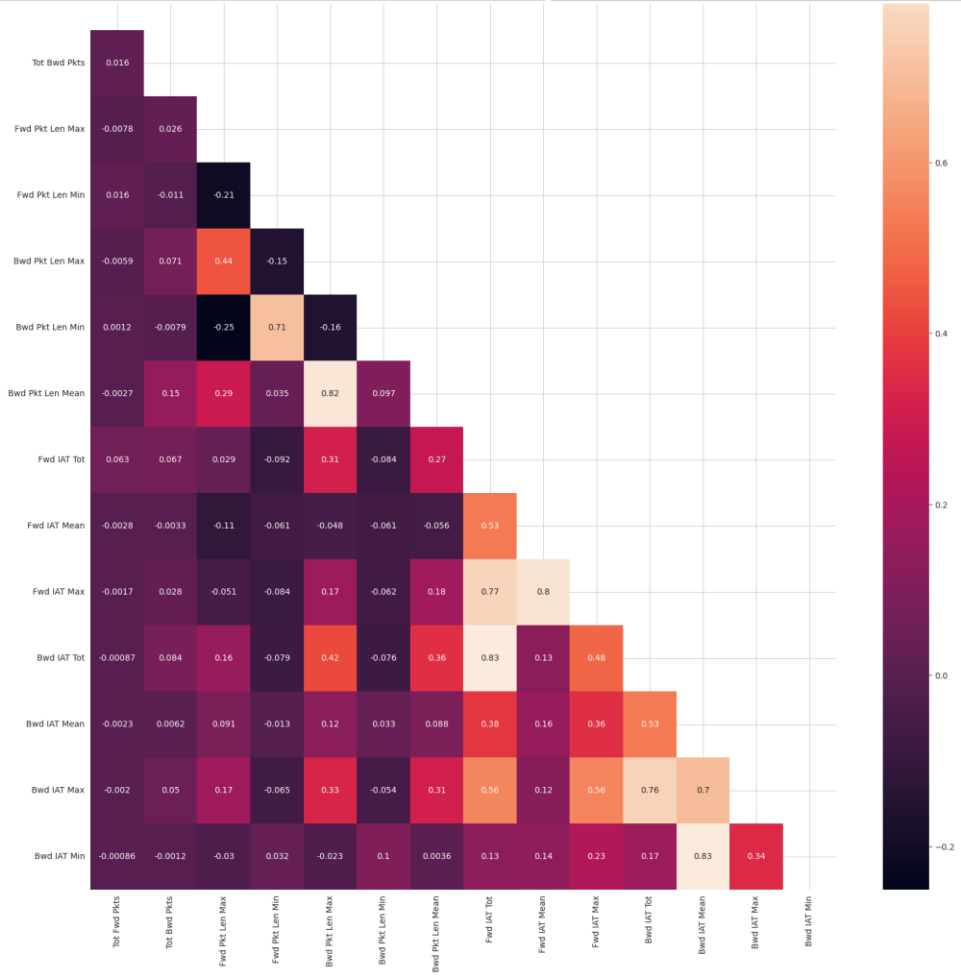
Data Sanitizing

Knowledge-based filtering → Correlation analysis (fwd) → Correlation analysis (bwd) → Final correlation analysis



Data Sanitizing

Knowledge-based filtering → Correlation analysis (fwd) → Correlation analysis (bwd) → Final correlation analysis



Selected Features (25/84)

```
data_clean.head()
```

	Flow Duration	Tot Fwd Pkts	Tot Bwd Pkts	Fwd Pkt Len Max	Fwd Pkt Len Min	Bwd Pkt Len Max	Bwd Pkt Len Min	Bwd Pkt Len Mean	Flow IAT Mean	Flow IAT Max	...	Bwd IAT Max	Bwd IAT Min	Pkt Len Min	Pkt Len Max	Pkt Len Mean	FIN Flag Cnt	SYN Flag Cnt	RST Flag Cnt	ACK Flag Cnt	Label
0	5032233	3	5	603.0	0.0	972.0	0.0	194.400000	7.188904e+05	5004303.0	...	5004614.0	6.0	0.0	972.0	175.000000	0	1	0	0	1
1	10993127	3	5	408.0	0.0	972.0	0.0	194.400000	1.570447e+06	5990709.0	...	5990709.0	3.0	0.0	972.0	153.333333	0	1	0	0	1
2	479425	3	7	142.0	0.0	1460.0	0.0	671.714286	5.326944e+04	232452.0	...	232857.0	227.0	0.0	1460.0	440.363636	0	1	0	0	1
3	9042123	1	1	0.0	0.0	0.0	0.0	0.000000	9.042123e+06	9042123.0	...	0.0	0.0	0.0	0.0	0.000000	0	0	0	1	1
4	1907579	1	1	0.0	0.0	0.0	0.0	0.000000	1.907579e+06	1907579.0	...	0.0	0.0	0.0	0.0	0.000000	0	0	0	1	1

5 rows × 26 columns

Data Pre-processing and Training

1. Data Preprocessing

Split dataset into:

- **Training set (75%)**
- **Testing set (25%)**

2. Model: Multi-Layer Perceptron (MLP)

Two hidden layers: **64 → 32 neurons**

Activation: **ReLU**

Optimizer: **Adam**

Training epochs: **200**

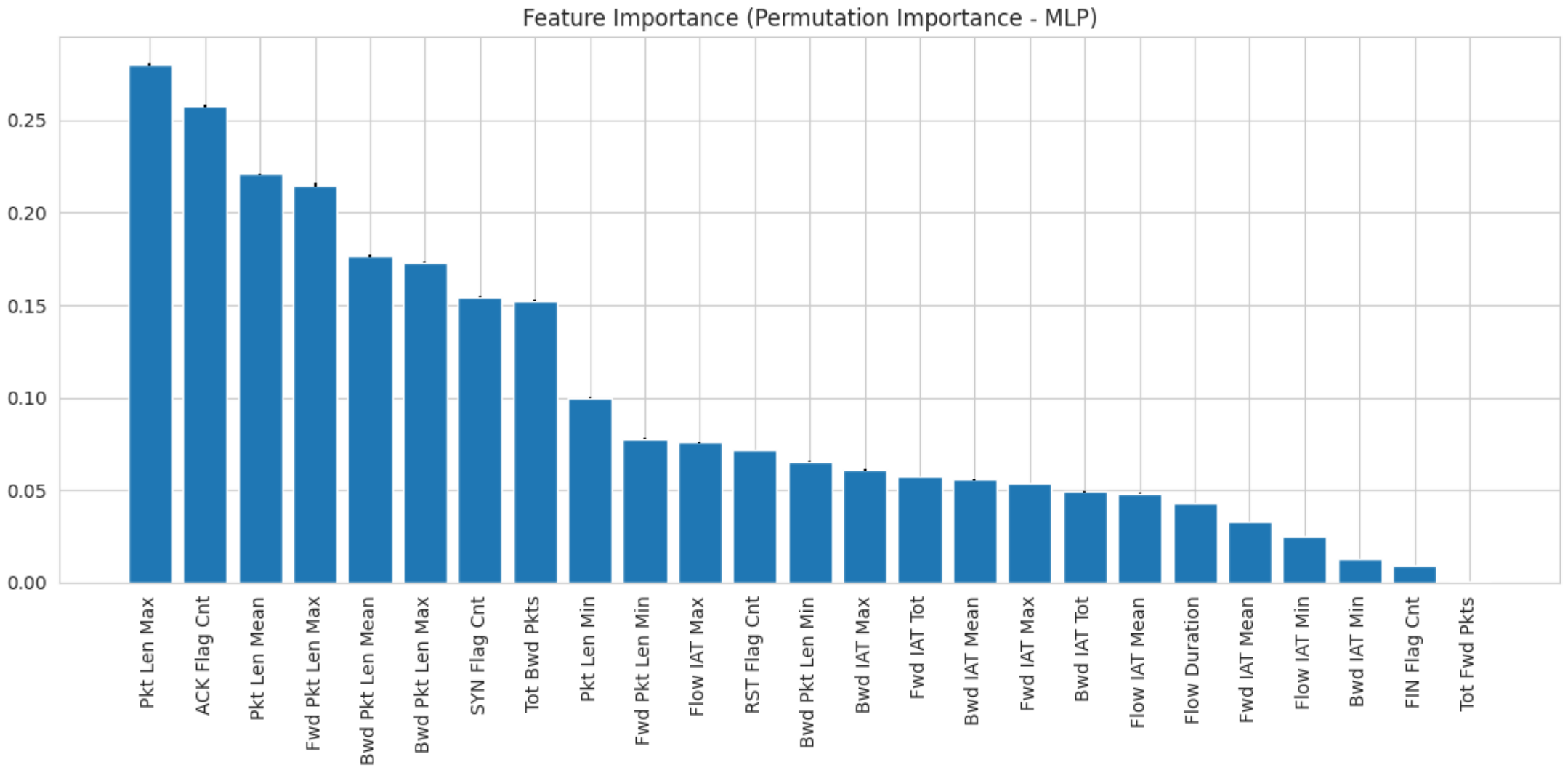
```
print(X_train.shape)
print(Y_train.shape)

print(X_test.shape)
print(Y_test.shape)
```

```
(959597, 25)
(959597, 2)
(319866, 25)
(319866, 2)
```

```
Accuracy: 0.9754365890716737

Classification Report:
              precision    recall  f1-score   support

   Benign 0       1.00      0.95      0.97    157931
    DDoS  1       0.96      1.00      0.98    161935

   micro avg       0.98      0.98      0.98    319866
   macro avg       0.98      0.98      0.98    319866
weighted avg       0.98      0.98      0.98    319866
 samples avg       0.98      0.98      0.98    319866
```

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Feature Importance



Future Steps

- Consider combinations of features (e.g., top 4-5 features)
- Re-train with the selected features.
- Testing more models for comparison
- Process and test with real packets from PCAP files
- Deploy in a real network environment
- Compare against work in the literature

THANK YOU!